東洋大学学術情報リポジトリ　Toyo University Repository for Academic Resources

機械・設計・哲学

著者	松浦 和也
著者別名	MATSUURA Kazuya
雑誌名	国際哲学研究
巻	別冊13
ページ	49-56
発行年	2020-03
URL	http://doi.org/10.34428/00011547

Creative Commons : 表示 - 非営利 - 改変禁止
http://creativecommons.org/licenses/by-nc-nd/3.0/deed.ja

機械・設計・哲学

松浦　和也

１．ロボットに関する哲学的問い

ロボットは心や意識を持つか、人工知能は自我を持つか、といった問いは、理学者や工学者も巻き込み、昨今頻繁に提起されている哲学的問題である。また、そもそも人工知能とは何か、という問いも、哲学的には有意義な問いだろう。というのは、この問いには２つのより根源的な問い、すなわち、人工物とは何か、という技術一般に関する問いと、知能とは何か、というより魂や精神と言った古来より論じられてきた問いを内包するからである。それゆえ、人工知能研究やロボット研究による成果は、哲学者や倫理学者が思索を深めるための良い素材を提供するだろう。

だが、ロボットや人工知能とは何か、というような定義の探求は、実際にその研究開発に携わる人々にとってはそれほど意義あるものではないかもしれない。ロボットや人工知能が心や意識を持つか、という問いも同様である。第一に、たしかに、ロボットや人工知能の定義は、もちろん非専門家にそれらを説明したり、これからロボットや人工知能の社会の中での扱いを法的に定めたりするために必要な手続きであるし、これらにまつわる倫理を論ずる時にも、議論対象の定義を判明にしておくことが求められる。しかし、そこで与えられた定義が実像と合わないからと言って、実際のロボットや人工知能に関する研究開発活動が大きく変容するものでもないだろう。そもそも、これらを研究開発する意義には、それまで考えられてきたロボットや人工知能の概念を越境することが含まれるはずである。実際に、現代の技術開発に法制度が追い付いていないという社会状況は、ロボットや人工知能の定義や概念自体が頻繁にアップデートされているという状況を反映している。第二に、研究開発者がいわゆるチューリング・テストに信頼を置き[1]、さらにそのテストを知能だけではなく、他の人間の能力にも拡張できると信じる限り、制作物に実際に（人間と同じようなシステムによる）心や意識を実装する必要はない。そうではなく、その制作物に心や意識があるように人々に見えさえすればよい。この立場に立つとき、ロボットや人工物は自我を持つかという問いは、果たして目の前の人間もこの私と同じような自我を持っているか、という超越論的問いを追いやるのと同時に、意味をなさないものとなる。

とはいえ、ロボットや人工知能が自我や意識を持つように振る舞うためにはどのようにすればよいか、という課題には興味を惹かれる研究開発者も少なくないだろう。特に、

コミュニケーションロボットやコンピューターゲームのキャラクターなどといった人間の代替や人間の操作を代行するものを研究開発する人にとっては、この課題は不可避のものである。このような場合、人間の心理メカニズムや神経伝達メカニズムをコンピューター上で模倣することは、ひとつの有力な手段となるだろう。もちろん、このような研究開発者たちの試みに対し、これまでの哲学的考察が提示してきた心理モデルや道徳発達モデルは示唆を与えることは可能である。実際に古典的な哲学的考察が提示したモデルが開発に利用される実例はすでに存在する[2]。

ただし、人間の代替となるようなものを制作することは、直ちに倫理的・道徳的問題を生じさせる。その局面で、ロボットや人工知能を道徳的主体（moral agent）とするにはどうすればよいか、という課題がしばしば検討される[3]。そこからは、人間のある特定の能力、たとえば感情や痛み、判断力、自由意志、自律性といったものを模倣することで、人工物に道徳的主体性を獲得させようとする試みがなされることになる[4]。

２．製品と道徳性

しかしながら、仮に道徳的主体であるための条件を満たしたロボットや人工知能が開発されたとして、その成果をどこまで社会の中で実用化できるだろうか。この問いに楽観的に応答することは難しい。

そのような自律的な機械を製品化したならば、たいていの場合、その機械は設計者と使用者の意図通りに動作するであろう。しかし、それらに不幸な事態や非倫理的な事態を引き起こす可能性が残っていることは想像に難くない。われわれの身の回りにあるカッターナイフや自動車が事故を起こしたり、時々犯罪に使われたりしていることを思い起こせば、同様の事態はこれら自律的な機械にも生じると想像することは自然である。そもそも、われわれの周囲にある製品で、事故を起こしえないものや、悪用できないようなものはあるだろうか[5]。子供心を思い出せば、たくさんの悪用の方法がありそうである。

もちろん、他の製品と同じく、自律的な機械を販売するメーカーは、欠陥がないように、誤用や悪用を防ぐように製品を設計することだろう。2020 年現在の日本では「製造物責任法」が施行されており、製造物に欠陥があった場合、消費者が被った被害はメーカーが保証することになっている。また、この法律が扱う対象外の製品であったり[6]、この法律の施行自体をやめたりしたとしても、消費者は欠陥がありそうな製品を避けるだろう。

では、道徳的主体であるように振る舞う自律機械は、メーカーにとっても消費者にとっても欠陥がない製品となり得るだろうか。この問いには、われわれが道徳的主体とみなされていることを自覚したうえで、われわれの振る舞いを反省すれば答えることができる。われわれ人間は常に道徳的に妥当であるように振る舞ってはおらず、それゆえ、ある動作主が道徳的主体であるための能力や機能を備えることと、その動作主が道徳的に振る舞うこととは異なる。したがって、自律機械が道徳的主体となるような機能を備えたとして

も、その機械は道徳的に問題のない動作をするわけではない。

もし、ロボットや人工知能を道徳的に振る舞わせようとするなら、別のシステムやアルゴリズム、あるいはデータが必要となる。しかし、現在の（筆者が理解する限りでの）技術の延長線上にありそうな試みはうまく成功するようには感じられない。たとえば、道徳的に振る舞うために従うべきルールをアルゴリズムに埋め込むことで機械の振る舞いの道徳性を確保しようと考える人もいるかもしれない。だが、われわれ人間はそれに従えば必ず道徳的であるような明文化されたルールを未だ獲得してはいない。それゆえ、何を機械に何を埋め込むかすら未だ判明ではない[7]。あるいは、人間の振る舞いを機械学習させることで機械に道徳的振る舞いを可能にさせるようと考える人もいるかもしれない。しかし、データを提供した人間たちの振る舞いが道徳的かどうかは疑念の余地がある。また、サンプルが少ないデータは強化学習に不向きである以上、例外的な局面での道徳的な振る舞いを機械学習によって獲得することも難しい。日常生活においてわれわれはたしかに嘘をつくべきではないが、カント的義務論からの主張に反して嘘をつくことがかえって道徳的と評価される局面もあるだろう。

しかし、消費者が期待し、メーカーが提供すべきは欠陥がない製品なのだとしたら、自律性を持つ機械に要求される道徳性とは、その製品がいかなる局面でも道徳的に振る舞うような、きわめて強い意味での完全な道徳性ではないだろうか。この要求の達成は、上述のように絶望的である。

さて、ある機械の振る舞いから生じた被害や不都合な帰結に対し、制作に関わった人間や法人に対する法的な保護や例外規定の有無に関わらず、一般の人々がその機械を制作したメーカーや設計者を責めることには十分な理由がある。そのメーカーや設計者はそのような自律機械を作ることも作らないこともできたからである。そうだとしたときに、なぜ道徳的主体となるような機械を作りたいと考えるのだろうか。偏屈な見方をすれば、その目的の裏側には、そのような機械を製造することによって、制作物の不都合が生じたとしても、自分自身にまで責めが及ぶことがないようにしたいという製作者側の密かな欲求があるようにすら感じられる。たしかに、ある人が制作したロボットが事故を起こしたとしても、そのロボットが道徳的主体であれば、人間が起こした事故に対する責めが彼／彼女の親に及ばないのと同様に、その製作者に責めを負わせることはないだろう。しかし、ここでわれわれは人間以外の道徳的主体を制作することは道徳的か、という別の倫理的問題に出会うことになる。

とはいえ、機械の振る舞いをすべて道徳的なものにすることも、道徳的主体であるために必要な能力を備えた自律機械を作成することも現状不可能なのであれば、また、道徳的主体であることと道徳的に振る舞うことは別物なのであれば、ロボットや人工知能を道徳的に問題が生じない、あるいは問題が生じる余地が少ない環境に置いたり、そのような環境を作り出そうしたりすることは理に適っている。たとえば、無人の工場では、（雇用の喪失といった社会的問題を除いて）、ロボットが直接人間に対し非道徳的振る舞いをす

ることはほとんどない。また、人間や人間が運転する乗り物は入れず、自動運転車しか走れない道路をインフラとして整備することは、道徳的問題の発生を小さくする効果があろう。自律機械のこのような社会実装の手段は、道徳的な自律機械を設計するというよりも、非道徳的ではないように環境の側を設計するという形で特徴づけられる。

３．無垢な機械と悪意

それでもなおロボットや人工知能は、もし円滑な形で社会に流通すれば、人間にこれまで以上の利便をもたらし、結果として人間を幸せにするだろうと信じられている。ここでは、この考えが妥当であるかは一度括弧にくくり、ロボットや人工知能がもたらす利便性を享受しつつも、社会に害悪にならないようにするための基本的な方向を探ることにしよう。

先に挙げた「製造物責任法」には免責条件がある。同法第四条第２項は、「当該製造物をその製造業者等が引き渡した時における科学又は技術に関する知見によっては、当該製造物にその欠陥があることを認識することができなかったこと」と定めている。ここで、「科学又は技術に関する知見」を参照しつつ、メーカーに製品の欠陥がないかを予め検証することを同法は求めている。

では、このような予見はどこまで可能であろうか。そして、予見できたからと言って、その不幸な帰結を防止することは可能だろうか。自動運転車を例にとろう。自動運転車の事故につながる原因は、人間がハンドルとアクセルで操作する自動車よりも膨大にありそうである。自動車に想定されるハードウェア的な不都合に加え、アルゴリズムのバグ、アップデートの失敗といったものも想定される。しかし、単純なアルゴリズムならともかく、複雑化した自動運転車のシステムからバグやアップデートの失敗を完全に取り除くことは困難である。のみならず、搭乗者の不明瞭な指示や過積載といった、どちらかといえば搭乗者に非があると判断されそうなものも、もしかしたらメーカーが予見すべきことに含まれるかもしれない。事故が起きた後で、不明瞭な指示や過積載を感知するようなシステムや機構を搭載すべきであった、という非難があってもおかしくはない。それだけではなく、もし遠距離操作を自動運転車に搭載するならば、外部からの操作の乗っ取りにも対応しなければならない。さらに、画像認識アルゴリズムの不都合をつけば、それを知る人間であれば事故を自動運転車に起こさせることもできるだろう。このように、事故や悪用につながる不都合すべてを予め列挙することは不可能である。また、不都合をとにかく生じさせないようにするために、数多くのメカニズムを単一の機械に搭載したとしても、それらを制御するためのアルゴリズムも複雑化する。

機械の中でも自律的な機械が引き起こすだろう事故が社会的に厄介なものとなりそうな理由の一つは、事故や不都合に関する予見を積み重ねれば積み重ねただけシステム全体の構造が複雑化することにより、結果として事故の原因を特定することが困難となる

ことである。もちろん、このことも予見可能性の範囲を法律や判例を通じて固定化し、他方で事故の原因を判別できるような技術的進歩を通じて解決可能かもしれない[8]。しかし、さらに真剣に取りざたされるべきは、自律機械を受け入れる環境、あるいは社会の側に含まれている。

悲しむべきことではあるが、われわれの社会には非道徳的行為を成す人間が存在する。のみならず、他者に非道徳的な行為を成すように命ずる人間も存在する。もちろん非道徳的行為を成した人間は非難されるだろう。では、非道徳的行為を成すように命じた人間は、実際にはその行為を成していないという理由で非難を免れるのだろうか。そのような免責は適切ではないだろう。なぜなら、そのような人は誰かを操ることで間接的にはいかなる非道徳的行為を成すことができることになるからである。

このような見方が妥当ならば、自律機械の社会実装が抱える別の倫理的問題が現れる。いわゆるアシモフのロボット三原則を遵守する人型ロボットが目の前にいたとしよう。このロボットは、第一条により[9]、人間に危害を加えることはない。ある人がロボットに人を殺すように命じたとしても、第二条が定める例外により[10]、その命令に従うことはない。しかし、このようなロボットに、直接的には無理であっても、殺人を間接的に指示する方法はいくらでもありそうである。すぐに思いつく方法は、被害者に布をかぶせたり、箱の中に入れたりして、ロボットのセンサーが人間とは認識できないようにすることである。その状態でそのロボットに「この包みをナイフで刺せ」、あるいは「この包みをロープできつく縛り、川に投げ捨てろ」と命令すれば、この命令の先にある悪意は達成される。それ以外にも、ロボットのセンサーを汚したり、壊したりすることでも実現が可能かもしれない。あるいは、卓越した技術者であれば、遠隔操作によるプログラムの書き換えをはじめとした様々な方法で、その操作の痕跡が残らないように、ロボットに実行させることができるだろう。

このように自律機械を悪用する方法は、原始的なものであれ、技術力を要求するものであれ、少し想像力を働かせさえすれば数多く想定できる。このような多様な悪用の方法が「その製造業者等が引き渡した時における科学又は技術に関する知見」に含まれるのであれば、メーカーは悪用の方法をできるだけたくさん想定したうえで、その方法を封じるように製品を設計せねばならないことになる。

しかし、社会に一定数存在する悪意を持った人間による命令や操作を受け付けないようなアルゴリズムを作ることはできるのだろうか。もし、それが可能であれば、完全な道徳性を備えた製品やロボットを作ることができるだけではなく、われわれ人間も完全に道徳的な振る舞いを常に行うことができるかもしれない。だが、その理想に到達することはできてはいない。それどころか、到達するための道すら明らかではない。

４．機械の倫理的設計に向けて

不都合や悪意ある使用法を予見して、それを全て防ぐような製品を設計することは原理的にできない、としよう。また、使用者の命令に悪意があった場合にその命令に従わないように、言い換えれば、道徳的判断を一任できるように、機械のアルゴリズムを設計することもできないとしよう。それにも関わらず、自律機械を社会の中で活躍させるためにはいかなる条件が必要か。

私見であるが、道徳的主体性を備えた自律機械をそのまま社会に流通させることは、人間の悪意を知らない無垢な幼児を社会の真っただ中に放り出すことに近似している。そうであるならば、そのような幼児は庇護の対象であるように、自律機械も庇護の対象として扱うべきなのではないか。ただし、機械は人間ではない。社会性と自身を守る術を長い時間をかけた教育によって身につける、という方法は採りようがない。

他方、悪戯や悪用によるものも含めた自律機械の事故に際し、メーカーと製作者を非難から遠ざけるための明確な条件ははっきりしている。人間や社会に悪しき影響を与える動作や出力を自律機械がしないことである。そのための一つの方法は、入力から出力に至るまでのプロセスの中で、悪い影響を与える出力が生成しないようにすればよい。ただし、善悪に関する予測をロボットや人工知能が原理的に困難だとしたら、最も確実で安全な方法はそのような機能を持たせないように自律機械を設計することである。前節のロボットを例にとれば、そもそもナイフを持ったり、ロープでものをくくることができたりするようなマニピュレータを装着させなければよい。もちろん、こうすることによってそのロボットは不便なものとなるだろう。しかし、それでも、そのロボットを社会の中で守ることはできる。

ただし、この提案はすべてに通用するようなものではないだろう。ロボットにマニピュレータを装着すべき場面も想定できるし[11]、もしかしたら、ロボットや人工知能の判断が有益かつ倫理的である可能性は残される[12]。

以上の考察から示唆されることは次のことである。第一に、あらゆる場面で人間の代替となるような「強い AI」とそれに従う身体を伴ったロボットは、社会の中で少なくとも製品としては流通させるに値しない。第二に、ロボットや人工知能を使用したり、人間が判断を委ねたりすることは有益かつ倫理的である余地は排除できないが、その場面や目的は限定せねばならない。第三に、製品としての自律機械が満たすべき倫理性、道徳性を個々の場面や目的に応じて判断すべきなのであれば、道徳的な自律機械の設計とはいかなるものか、または少なくとも非道徳的ではない自律機械の設計とはいかなるものかについて、生産的な知見を提供するのは哲学や倫理学をはじめとした人文学であろう。

*本稿は RISTEX、JPMJRX17H3、JSPS 課題研究設定による先導的人文学・社会科学研究推進事業 JSPS001 18070707 の委託を受けたものです。

註

1 cf. Turing, 433-434.

2 たとえば、三宅が示すように、コンピューターゲームの開発に哲学的考察は用いられている。

3 その種の検討の見取り図として、Allen は有益である。

4 たとえは、浅田は、痛覚をロボットに実装させることから、道徳的行為者へ進展させる作業仮説を提示している。浅田, p. 19.

5 ここで、悪用されることを拒否するといったアルゴリズムを作ればよい、と提案されるかもしれない。だが、何が悪用で何が悪用でないかを普遍的に判断できるような道徳的基準をわれわれは持ち合わせているだろうか。

6 「製造物責任法」第二条には、「製造物」を動産に制限している。プログラムやデータ、アルゴリズムは動産ではないので、欠陥があったとしても直ちにその種の製品に同法が適用されるわけではないだろう。

7 このような機械の倫理性に関するトップダウン的アプローチとその問題については、ウォラック, pp. 113-134 参照。

8 ただし、現在の証券市場が望ましくない値動きをした場合に、原因を探ることはどこまで可能か。複数の自動売買アルゴリズムはそれぞれひとつのシステムであるが、それらが複合してひとつの「システム」を形成している。それぞれのアルゴリズムは完全に動作しているにも関わらず、あるいはむしろ完全に動作していることによって、実態に合わない株価の暴落といった社会に不利益をもたらすことはすでに生じている。

9 「ロボットは人間を傷つけてはならない。あるいは、動作しないことによって、人間が害されることになることを無視してはならない。」

10 「ロボットは人間から与えられた命令に従わねばならない。ただし、その命令が第一条に反する場合を除く。」

11 病院等の福祉のために導入されるロボットにはそのような機能が求められよう。

12 たとえば、精神的に障害がある人々の判断をサポートするような人工知能は多くの場合有益であるかもしれない。

【参考文献】

Allen, C. Varner, G. ‘Prolegomena to any future artificial moral agent’. *Journal of Experimental & Theoretical Artificial Intelligence,* 12(3), pp. 251-261. 2000.

浅田稔「人工痛覚が導く意識の発達過程としての共感、モラル、倫理」『哲学』70, pp. 14-34. 2019.

三宅陽一郎『人工知能のための哲学塾』ビー・エヌ・エヌ新社, 2016.

Nagenborg, M. et al. ‘Ethical regulations on robotics in Europe’. *AI & Society*. 22, 3, pp. 349-366. 2008.

Turing, A. M. ‘Computing Machinery and Intelligence’. *Mind*. LIX, 236, pp. 433-460. 1950.